Building Proteins in a Day: Efficient 3D Molecular Reconstruction

Marcus A. Brubaker Ali Punjani David J. Fleet
University of Toronto
{mbrubake,alipunjani,fleet}@cs.toronto.edu

Abstract

Discovering the 3D atomic structure of molecules such as proteins and viruses is a fundamental research problem in biology and medicine. Electron Cryomicroscopy (Cryo-EM) is a promising vision-based technique for structure estimation which attempts to reconstruct 3D structures from 2D images. This paper addresses the challenging problem of 3D reconstruction from 2D Cryo-EM images. A new framework for estimation is introduced which relies on modern stochastic optimization techniques to scale to large datasets. We also introduce a novel technique which reduces the cost of evaluating the objective function during optimization by over five orders or magnitude. The net result is an approach capable of estimating 3D molecular structure from large scale datasets in about a day on a single workstation.

1. Introduction

Discovering the 3D atomic structure of molecules such as proteins and viruses is a fundamental research problem in biology and medicine. The ability to routinely determine the 3D structure of such molecules would potentially revolutionize the process of drug development and accelerate research into fundamental biological processes. Electron Cryomicroscopy (Cryo-EM) is an emerging vision-based approach to 3D macromolecular structure determination that is applicable to medium to large-sized molecules in their native state. This is in contrast to X-ray crystallography which requires a crystal of the target molecule, which are often impossible to grow [32] or nuclear magnetic resonance (NMR) spectroscopy which is limited to relatively small molecules [15].

The Cryo-EM reconstruction task is to estimate the 3D density of a target molecule from a large set of images of the molecule (called particle images). The problem is similar in spirit to multi-view scene carving [6, 16] and to large-scale, uncalibrated multi-view reconstruction [1]. Like multi-view scene carving, the goal is to estimate a dense 3D occupancy representation of shape from a set of different views, but

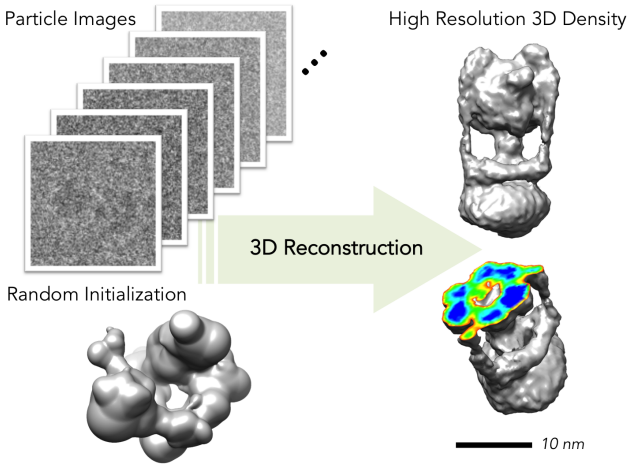


Figure 1: The goal is to reconstruct the 3D structure of a molecule (right), at nanometer scales, from a large number of noisy, uncalibrated 2D projections obtained from cryogenically frozen samples in an electron microscope (left).

unlike many approaches to scene carving, we do not assume calibrated cameras, since we do not know the relative 3D poses of the molecule in different images. Like uncalibrated, multi-view reconstruction, we aim to use very large numbers of uncalibrated views to obtain high fidelity 3D reconstructions, but the low signal-to-noise levels in Cryo-EM (often as low as 0.05 [4]; see Fig. 1) and the different image formation model prevent the use of common feature matching techniques to establish correspondences. Computed Tomography (CT) [13, 10] uses a similar imaging model (orthographic integral projection), however in CT the projection direction of each image is known whereas with Cryo-EM the relative pose of each particle is unknown.

Existing Cryo-EM techniques, *e.g.*, [7, 11, 34, 37], suffer from two key problems. First, without good initialization, they converge to poor or incorrect solutions [12], often with little indication that something went wrong. Second, they are extremely slow, which limits the number of particles images one can use as input to mitigate the effects of noise; *e.g.*, the website of the RELION package [34] reports

requiring two weeks on 300 cores to process a dataset with 200,000 images.

We introduce a framework for Cryo-EM density estimation, formulating the problem as one of stochastic optimization to perform maximum-a-posteriori (MAP) estimation in a probabilistic model. The approach is remarkably efficient, providing useful low resolution density estimates in an hour. We also show that our stochastic optimization technique is insensitive to initialization, allowing the use of random initializations. We further introduce a novel importance sampling scheme that dramatically reduces the computational costs associated with high resolution reconstruction. This leads to speedups of 100,000-fold or more, allowing structures to be determined in a day on a modern workstation. In addition, the proposed framework is flexible, allowing parts of the model to be changed and improved without impacting the estimation; *e.g.*, we compare the use of three different priors. To demonstrate our method, we perform reconstructions on two real datasets and one synthetic dataset.

2. Background and Related Work

In Cryo-EM, a purified solution of the target molecule is cryogenically frozen into a thin (single molecule thick) film, and imaged with a transmission electron microscope. A large number of such samples are obtained, each of which provides a micrograph containing hundreds of visible, individual molecules. In a process known as *particle picking*, individual molecules are selected, resulting in a stack of cropped *particle images*. Particle picking is often done manually, however there have been recent moves to partially or fully automate the process [17, 40]. Each particle image provides a noisy view of the molecule, but with unknown 3D pose, see Fig. 2 (right). The reconstruction task is to estimate the 3D electron density of the target molecule from the potentially large set of particle images.

Common approaches to Cryo-EM density estimation, *e.g.*, [7, 11, 37], use a form of iterative refinement. Based on an initial estimate of the 3D density, they determine the best matching pose for each particle image. A new density estimate is then constructed using the Fourier Slice Theorem (FST); *i.e.*, the 2D Fourier transform of an integral projection of the density corresponds to a slice through the origin of the 3D Fourier transform of that density, in a plane perpendicular to the projection direction [13]. Using the 3D pose for each particle image, the new density is found through interpolation and averaging of the observed particle images.

This approach is fundamentally limited in several ways. Even if one begins with the correct 3D density, the low SNR of particle images makes accurately identifying the correct pose for each particle nearly impossible. This problem is exacerbated when the initial density is inaccurate. Poor initializations result in estimated structures that are either clearly wrong (see Fig. 9) or, worse, appear plausible but are misleading in reality, resulting in incorrectly estimated 3D structures [12]. Finally, and crucially for the case of density estimation with many particle images, all data are used at each refinement iteration, causing these methods to be extremely slow. Mallick et al. [25] proposed an approach which attempted to establish weak constraints on the relative 3D poses between different particle images. This was used to initialize an iterative refinement algorithm to produce a final reconstruction. In contrast, our refinement approach does not require an accurate initialization.

To avoid the need to estimate a single 3D pose for each particle image, Bayesian approaches have been proposed in which the 3D poses for the particle images are treated as latent variables, and then marginalized out numerically. This approach was originally proposed by Sigworth [35] for 2D image alignment and later by Scheres et al. [33] for 3D estimation and classification. It was since been used by Jaitly et al. [14], where batch, gradient-based optimization was performed. Nevertheless, due to the computational cost of marginalization, the method was only applied to small numbers of class-average images which are produced by clustering, aligning and averaging individual particle images according to their appearance, to reduce noise and the number of particle images used during the optimization. More recently, pose marginalization was applied directly with particle images, using a batch Expectation-Maximization algorithm in the RELION package [34]. However, this approach is extremely computationally expensive. Here, the proposed approach uses a similar marginalized likelihood, however unlike previous methods, stochastic rather than batch optimization is used. We show that this allows for efficient optimization, and for robustness with respect to initialization. We further introduce a novel importance sampling technique that dramatically reduces the computational cost of the marginalization when working at high resolutions.

3. A Framework for 3D Density Estimation

Here we present our framework for density estimation which includes a probabilistic generative model of image formation, stochastic optimization to cope with large-scale datasets, and importance sampling to efficiently marginalize over the unknown 3D pose of the particle in each image.

3.1. Image Formation Model

In Cryo-EM, particle images are formed as orthographic, integral projections of the electron density of a molecule, $\mathcal{V} \in \mathbb{R}^{D^3}$. In each image, the density is oriented in an unknown pose, $\mathbf{R} \in \mathcal{SO}(3)$, relative to the direction of the microscope beam. The projection along this unknown direction is a linear operator, which is represented by the matrix $\mathbf{P}_{\mathbf{R}} \in \mathbb{R}^{D^2 \times D^3}$. Along with pose, the in-plane translation $\mathbf{t} \in \mathbb{R}^2$ of each particle image is unknown, the effect of

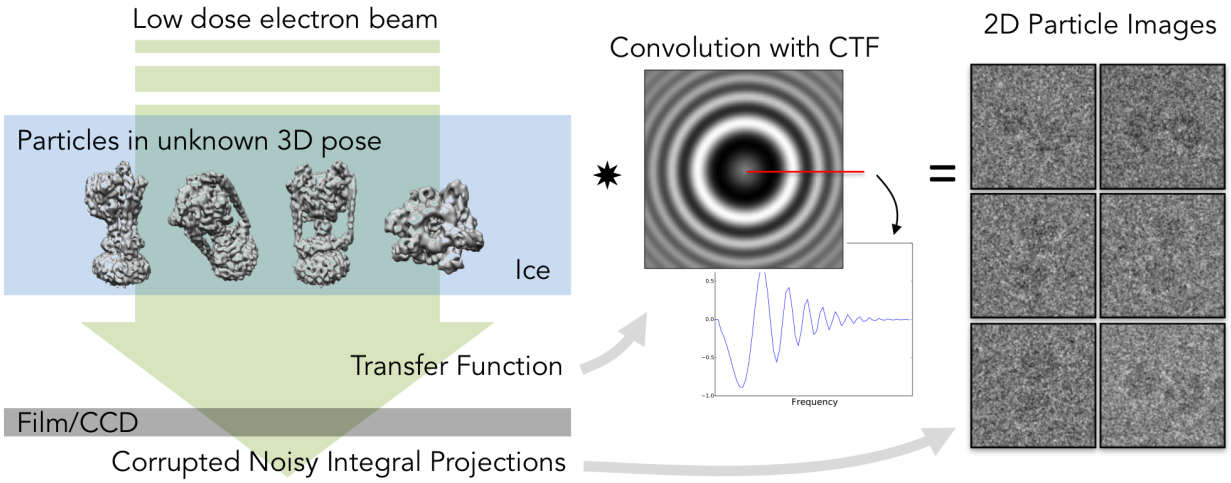


Figure 2: A generative image formation model in Cryo-EM. The electron beam results in an orthographic integral projection of the electron density of the specimen. This projection is modulated by the Contrast Transfer Function (CTF) and corrupted with noise. The images pictured here showcase the low SNR typical in Cryo-EM. The zeros in the CTF (which completely destroy some spatial information) make estimation particularly challenging, however their locations vary as a function of microscope parameters. These are set differently across particle images in order to mitigate this problem. Particle images and density from [18].

which is similarly represented by a matrix $\mathbf{S_t} \in \mathbb{R}^{D^2 \times D^2}$. The resulting shifted projection is corrupted by two phenomena: a contrast transfer function (CTF) and noise. The CTF is analogous to the effects of defocus in a conventional light camera and can be modelled as a convolution of the projected image. This linear operation is represented here by the matrix $\mathbf{C}_\theta \in \mathbb{R}^{D^2 \times D^2}$ where θ are the parameters of the CTF model [30]. The Fourier spectrum of a typical CTF is shown in Figure 2; note the phase changes which result in zero crossings (not typically observed in traditional light cameras) and the attenuation at higher frequencies which makes estimation particularly challenging. CTF parameters, θ, are assumed to be given; CTF estimation is beyond the scope of this work, but is routinely done using existing tools, *e.g.*, [24, 27].

As noted above, and clearly seen in Figure 2, there is a large amount of noise present in typical particle images. This is primarily due to the sensitive nature of biological specimens, requiring extremely low exposures. The noise is modelled using an IID Gaussian distribution, resulting in the following expression for the conditional distribution of a particle image, $\mathcal{I} \in \mathbb{R}^{D^2}$,

$$p(\mathcal{I} \,|\, \theta, \mathbf{R}, \mathbf{t}, \mathcal{V}) = \mathcal{N}(\mathcal{I} \,|\, \mathbf{C}_\theta \mathbf{S_t} \mathbf{P_R} \mathcal{V}, \sigma^2 \mathbf{I}) \qquad (1)$$

where σ is the standard deviation of the noise and $\mathcal{N}(\cdot|\mu, \Sigma)$ is the multivariate normal distribution with mean vector μ and covariance matrix Σ.

In practice, due to computational considerations, Equation (1) is evaluated in Fourier space, making use of the Fourier Slice Theorem and Parseval's Theorem to obtain

$$p(\tilde{\mathcal{I}} \,|\, \theta, \mathbf{R}, \mathbf{t}, \tilde{\mathcal{V}}) = \mathcal{N}(\tilde{\mathcal{I}} \,|\, \tilde{\mathbf{C}}_\theta \tilde{\mathbf{S}}_\mathbf{t} \tilde{\mathbf{P}}_\mathbf{R} \tilde{\mathcal{V}}, \sigma^2 \mathbf{I}) \qquad (2)$$

where $\tilde{\mathcal{I}}$ is the 2D Fourier transform of the image, $\tilde{\mathbf{S}}_\mathbf{t}$ is the shift operator in Fourier space (a phase change), $\tilde{\mathbf{C}}_\theta$ is the CTF modulation in Fourier space (a diagonal operator), $\tilde{\mathbf{P}}_\mathbf{R}$ is a sinc interpolation operator which extracts a plane through the origin defined by the projection orientation $\mathbf{R}$ and $\tilde{\mathcal{V}}$ is the 3D Fourier transform of $\mathcal{V}$. To speed the computation of the likelihood, and due to the level of noise and attenuation of high frequencies by the CTF, a maximum frequency is specified, ω, beyond which frequencies are ignored.

The 3D pose, $\mathbf{R}$, and shift, $\mathbf{t}$, of each particle image are unknown and treated as latent variables which are marginalized out [35, 33]. Assuming $\mathbf{R}$ and $\mathbf{t}$ are independent of each other and the density $\mathcal{V}$, one obtains

$$p(\tilde{\mathcal{I}} \,|\, \theta, \tilde{\mathcal{V}}) = \int_{\mathbb{R}^2} \int_{\mathcal{SO}(3)} p(\tilde{\mathcal{I}}|\theta, \mathbf{R}, \mathbf{t}, \tilde{\mathcal{V}}) p(\mathbf{R}) p(\mathbf{t}) d\mathbf{R} d\mathbf{t} \qquad (3)$$

where $p(\mathbf{R})$ is a prior over 3D poses, $\mathbf{R} \in \mathcal{SO}(3)$, and $p(\mathbf{t})$ is a prior over translations, $\mathbf{t} \in \mathbb{R}^2$. In general, nothing is known about the projection direction so $p(\mathbf{R})$ is assumed to be a uniform distribution. Particles are picked to be close to the center of each image, so $p(\mathbf{t})$ is chosen to be a Gaussian distribution centered in the image. The above double integral is not analytically tractable, so numerical quadrature is used [22, 9]. The conditional probability of an image

(likelihood) then becomes

$$p(\tilde{\mathcal{I}}|\theta,\tilde{\mathcal{V}}) \approx \sum_{j=1}^{M_{\mathbf{R}}} w_j^{\mathbf{R}} \sum_{\ell=1}^{M_{\mathbf{t}}} w_k^{\mathbf{t}} p(\tilde{\mathcal{I}}|\theta,\mathbf{R}_j,\mathbf{t}_\ell,\tilde{\mathcal{V}}) p(\mathbf{R}) p(\mathbf{t}) \tag{4}$$

where $\{(\mathbf{R}_j, w_j^{\mathbf{R}})\}_{j=1}^{M_{\mathbf{R}}}$ are weighted quadrature points over $\mathcal{SO}(3)$ and $\{(\mathbf{t}_\ell, w_\ell^{\mathbf{t}})\}_{\ell=1}^{M_{\mathbf{t}}}$ are weighted quadrature points over $\mathbb{R}^2$. The accuracy of the quadrature scheme, and consequently the values of $M_{\mathbf{R}}$ and $M_{\mathbf{t}}$, are set automatically based on ω, the specified maximum frequency such that higher values of ω results in more quadrature points.

Given a set of K images with CTF parameters $\mathfrak{D} = \{(\mathcal{I}_i, \theta_i)\}_{i=1}^K$ and assuming conditional independence of the images, the posterior probability of a density $\mathcal{V}$ is

$$p(\mathcal{V}|\mathfrak{D}) \propto p(\mathcal{V}) \prod_{i=1}^{K} p(\tilde{\mathcal{I}}_i|\theta_i,\tilde{\mathcal{V}}) \tag{5}$$

where $p(\mathcal{V})$ is a prior over 3D molecular electron densities. Several choices of prior are explored below, but we found that a simple independent exponential prior worked well. Specifically, $p(\mathcal{V}) = \prod_{i=1}^{D^3} \lambda e^{-\lambda \mathcal{V}_i}$ where $\mathcal{V}_i$ is the value of the ith voxel and λ is the inverse scale parameter. Other choices of prior are possible and is a promising direction for future research.

Estimating the density now corresponds to finding $\mathcal{V}$ which maximizes Equation (5). Taking the negative log and dropping constant factors, the optimization problem becomes $\arg\min_{\mathcal{V}\in\mathbb{R}_+^{D^3}} f(\mathcal{V})$,

$$f(\mathcal{V}) = -\log p(\mathcal{V}) - \sum_{i=1}^{K} \log p(\tilde{\mathcal{I}}_i|\theta_i,\tilde{\mathcal{V}}) \tag{6}$$

where $\mathcal{V}$ is restricted to be positive (negative density is physically unrealistic). Optimizing Eq. (6) directly is costly due to the marginalization in Eq. (4) as well as the large number (K) of particle images in a typical dataset. To deal with these challenges, the following sections propose the use of two techniques, namely, stochastic optimization and importance sampling.

3.2. Stochastic Optimization

In order to efficiently cope with the large number of particle images in a typical dataset, we propose the use of stochastic optimization methods. Stochastic optimization methods exploit the large amount of redundancy in most datasets by only considering subsets of data (*i.e.*, images) at each iteration by rewriting the objective as $f(\mathcal{V}) = \sum_k f_k(\mathcal{V})$ where each $f_k(\mathcal{V})$ evaluates a subset of data. This allows for fast progress to be made before a batch optimization algorithm would be able to take a single step.

There are a wide range of such methods, ranging from simple stochastic gradient descent with momentum [28, 29, 36] to more complex methods such as Natural Gradient methods [2, 3, 19, 20] and Hessian-free optimization [26]. Here we propose the use of Stochastic Average Gradient Descent (SAGD) [21] which has several important advantages. First, it is effectively self-tuning, using a line-search to determine and adapt the learning rate. This is particularly important, as many methods require significant manual tuning for new objective functions and, potentially, each new dataset. Further, it is specifically designed for the finite dataset case allowing for faster convergence.

At each iteration τ, SAGD [21] considers only a single subset of data, k_τ, which defines part of the objective function $f_{k_\tau}(\mathcal{V})$ and its gradient $\mathbf{g}_{k_\tau}(\mathcal{V})$. The density $\mathcal{V}$ is then updated as

$$\mathcal{V}_{\tau+1} = \mathcal{V}_\tau - \frac{\epsilon}{KL} \sum_{j=1}^{K} d\mathcal{V}_j^\tau \tag{7}$$

where ϵ is a base learning rate, L is a Lipschitz constant of $\mathbf{g}_k(\mathcal{V})$, and

$$d\mathcal{V}_k^\tau = \begin{cases} \mathbf{g}_k(\mathcal{V}_\tau) & k = k_\tau \\ d\mathcal{V}_k^{\tau-1} & \text{otherwise} \end{cases} \tag{8}$$

is the most recent gradient evaluation of datapoint j at iteration τ. This step can be computed efficiently by storing the gradient of each observation and updating a running sum each time a new gradient is seen. The Lipschitz constant L is not generally known but can be estimated using a line-search technique. Theoretically, convergence occurs for values of $\epsilon \leq \frac{1}{16}$ [21], however in practice larger values at early iterations can be beneficial, thus we use $\epsilon = \max(\frac{1}{16}, 2^{1-\lfloor \tau/150 \rfloor})$. To allow parallelization and reduce the memory requires of SAGD, the data is divided into minibatches of 200 particles images. Finally, to enforce the positivity of density, negative values of $\mathcal{V}$ are truncated to zero after each iteration. More details of the stochastic optimization can be found in the Supplemental Material.

3.3. Importance Sampling

While stochastic optimization allows us to scale to large datasets, the cost of computing the required gradient for each image remains high due to the marginalization over orientations and shifts. Intuitively, one could consider randomly selecting a subset of the terms in Eq. (4) and using this as an approximation. This idea is formalized by importance sampling (IS) which allows for an efficient and accurate approximation of the discrete sums in Eq. (4).[1] A full review of importance sampling is beyond the scope of this paper but we refer readers to [38].

[1] One can also apply importance sampling directly to the continuous integrals in Eq. (3) but it can be computationally advantageous to precompute a fixed set of projection and shift matrices, $\tilde{\mathbf{P}}_{\mathbf{R}}$ and $\tilde{\mathbf{S}}_{\mathbf{t}}$, which can be reused across particle images.

To apply importance sampling, consider the inner sum from Eq. (4), rewriting it as

$$\phi_j^{\mathbf{R}} = \sum_{\ell=1}^{M_{\mathbf{t}}} w_\ell^{\mathbf{t}} p_{j,\ell} = \sum_{\ell=1}^{M_{\mathbf{t}}} q_\ell^{\mathbf{t}} \left(\frac{w_\ell^{\mathbf{t}} p_{j,\ell}}{q_\ell^{\mathbf{t}}} \right) \tag{9}$$

where $p_{j,\ell} = p(\tilde{\mathcal{I}}|\theta, \mathbf{R}_j, \mathbf{t}_\ell, \tilde{\mathcal{V}}) p(\mathbf{R}_j) p(\mathbf{t}_\ell)$ and $\mathbf{q}^{\mathbf{t}} = (q_1^{\mathbf{t}}, \ldots, q_{M_{\mathbf{t}}}^{\mathbf{t}})^T$ is the parameter vector of a multinomial importance distribution such that $\sum_{\ell=1}^{M_{\mathbf{t}}} q_\ell^{\mathbf{t}} = 1$ and $q_\ell^{\mathbf{t}} > 0$. The domain of $\mathbf{q}^{\mathbf{t}}$ corresponds to the set of quadrature points in Equation (4). Then, $\phi_j^{\mathbf{R}}$ can be thought of as the expected value $E_\ell[\frac{w_\ell^{\mathbf{t}} p_{j,\ell}}{q_\ell^{\mathbf{t}}}]$ where ℓ is a random variable distributed according to $\mathbf{q}^{\mathbf{t}}$. If a set of $N_{\mathbf{t}} \ll M_{\mathbf{t}}$ random indexes $\mathfrak{I}^{\mathbf{t}}$ are drawn according to $\mathbf{q}^{\mathbf{t}}$, then

$$\phi_j^{\mathbf{R}} \approx \frac{1}{N_{\mathbf{t}}} \sum_{\ell \in \mathfrak{I}^{\mathbf{t}}} \frac{w_\ell^{\mathbf{t}} p_{j,\ell}}{q_\ell^{\mathbf{t}}} \,. \tag{10}$$

Thus, we can efficiently approximate $\phi_j^{\mathbf{R}}$ by drawing samples according to the importance distribution $\mathbf{q}^{\mathbf{R}}$ and computing the average. Using this approximation in Eq. (4) gives

$$p(\tilde{\mathcal{I}}|\theta, \tilde{\mathcal{V}}) \approx \sum_{j=1}^{M_{\mathbf{R}}} w_j^{\mathbf{R}} \frac{1}{N_{\mathbf{t}}} \left(\sum_{\ell \in \mathfrak{I}^{\mathbf{t}}} \frac{w_\ell^{\mathbf{t}} p_{j,\ell}}{q_\ell^{\mathbf{t}}} \right) \tag{11}$$

and importance sampling can be similarly used for the outer summation to give

$$p(\tilde{\mathcal{I}}|\theta, \tilde{\mathcal{V}}) \approx \sum_{j \in \mathfrak{I}^{\mathbf{R}}} \frac{w_j^{\mathbf{R}}}{N_{\mathbf{R}} q_j^{\mathbf{R}}} \left(\sum_{\ell \in \mathfrak{I}^{\mathbf{t}}} \frac{w_\ell^{\mathbf{t}}}{N_{\mathbf{t}} q_\ell^{\mathbf{t}}} p_{j,\ell} \right) \tag{12}$$

where $\mathfrak{I}^{\mathbf{R}}$ are samples drawn from the importance distribution $\mathbf{q}^{\mathbf{R}} = (q_1^{\mathbf{R}}, \ldots, q_{M_{\mathbf{R}}}^{\mathbf{R}})^T$ used for approximating

$$\phi_\ell^{\mathbf{t}} = \sum_{j=1}^{M_{\mathbf{R}}} w_j^{\mathbf{R}} p_{j,\ell} \approx \frac{1}{N_{\mathbf{R}}} \sum_{j \in \mathfrak{I}^{\mathbf{R}}} \frac{w_j^{\mathbf{R}} p_{j,\ell}}{q_j^{\mathbf{R}}} \,. \tag{13}$$

The accuracy of the approximation in Eq. (12) is controlled by the number of samples used, with the error going to zero as N increases. We use $N = s_0 s(\mathbf{q})$ samples where $s(\mathbf{q}) = \left(\sum_\ell q_\ell^2\right)^{-1}$ is the effective sample size [8] and s_0 is a scaling factor. This choice ensures that when the importance distribution is diffuse, more samples are used.

While the estimates provided by IS are unbiased, their error can be arbitrarily bad if the importance distribution is not well chosen. To choose a suitable importance distribution, we make two observations. First, the values $\phi_\ell^{\mathbf{t}}$ and $\phi_j^{\mathbf{R}}$ are proportional to the marginal probability of single particle image having been generated with shift $\mathbf{t}_\ell$ or pose $\mathbf{R}_j$, making them natural choices on which to base the importance distributions. Second, these values remain stable once the rough shape of the structure has been determined. This can be seen in Figure 3 which shows that by the third epoch the KL divergence of the values of $\phi^{\mathbf{R}}$ from one epoch to the next is extremely small.

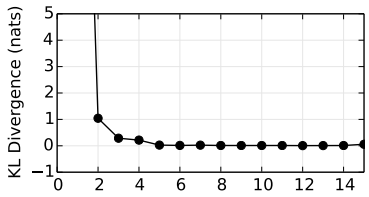
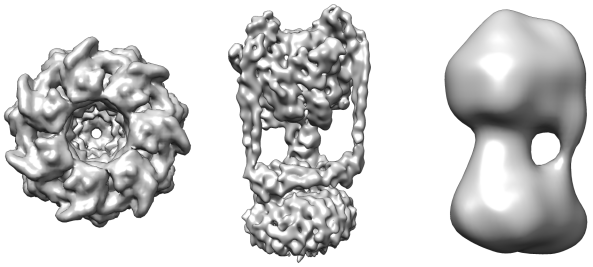


Figure 3: The KL divergence between the values of $\phi^{\mathbf{R}}$ at the current and previous epochs on the thermus dataset.



GroEL-GroES [39] Thermus ATPase [18] Bovine ATPase [31]

Figure 4: Previously published structures for the datasets used in this paper.

Thus we use these quantities, computed from the previous iterations, to construct the importance distributions at the current iteration. Dropping the $\mathbf{R}$ or $\mathbf{t}$ superscripts for clarity, let $\mathfrak{I}$ be the set of samples evaluated at the previous iteration and ϕ_i be the computed values for $i \in \mathfrak{I}$. Then the importance distribution used at the current iteration is

$$q_j = (1 - \alpha) Z^{-1} \hat{\phi}_j + \alpha \psi_j \tag{14}$$

where ψ_j is a uniform prior distribution, α controls how much the previous distribution is relied on, $Z = \sum_j \hat{\phi}_j$, and $\hat{\phi}_j = \sum_{i \in \mathfrak{I}} \phi_i^{1/T} \mathbf{K}_{i,j}$. Here T is an annealing parameter and $\mathbf{K}_{i,j}$ are entries of a kernel matrix computed on the quadrature points which diffuses probability to nearby quadrature points. The values for $\alpha = \max(0.05, 2^{-0.25 \lfloor \tau_{prev}/50 \rfloor})$ and $T = \max(1.25, 2^{10.0/\lfloor \tau_{prev}/50 \rfloor})$ are set so that at early iterations, when the underlying density is changing, we rely more heavily on the prior. For $\mathbf{K}$ we use a Gaussian kernel for the shifts and a Fisher kernel for the orientations. The bandwidth of the kernel is tuned based on the current resolution of the quadrature scheme, *e.g.*, the Gaussian shift kernel bandwidth is set to be equal to the spacing between the shift quadrature points. More details on importance sampling can be found in the Supplemental Material.

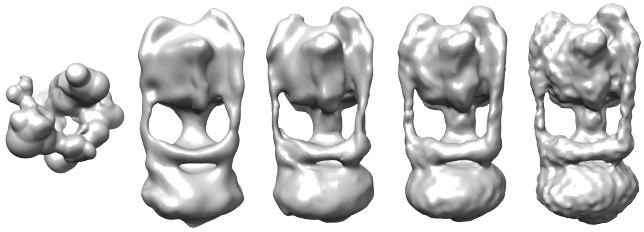


Figure 5: The random initialization (left) used in all experiments, generated by summing random spheres, and reconstruction of the *thermus* dataset after various amounts of computation. Note that within an hour of computation, the gross structure is already well determined, after which fine details emerge gradually.

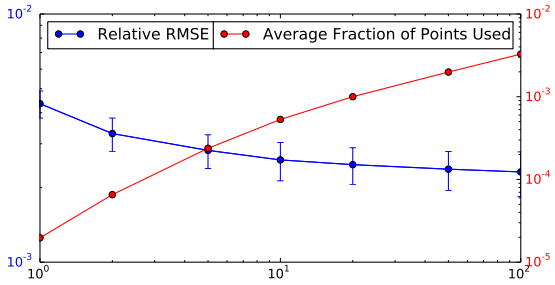


Figure 6: Relative error (blue, left axis) and fraction of total quadrature points (red, right axis) used in computing $\log p(\tilde{\mathcal{I}}|\theta, \tilde{\mathcal{V}})$ as a function of the ESS scaling factor, s_0 (horizontal axis). Note the log-scale of the axes.

4. Experiments

The proposed method was applied to two experimental datasets and one synthetic dataset. All experiments used the same parameters and were initialized using the same randomly generated density shown in Figure 5(left). The maximum frequency considered was gradually increased from an initial value of $\omega = 1/40$Å to a maximum of $\omega = 1/10$Å. This maximum frequency corresponds to the resolution of the best published results for the datasets used here, *i.e.*, [18]. Optimizations were run until the maximum resolution was reached and the average error on a held-out set of 100 particle images stopped improving, around 5000 iterations.

Datasets The first dataset was ATP synthase from the *thermus thermophilus* bacteria, a large transmembrane molecule. The *thermus* dataset consisted of $46,105$ particle images which were provided by Lau and Rubinstein [18]. The high resolution structure from [18] and some sample images are shown in Figure 2. The second dataset was *bovine* mitochondrial ATP synthase [31]. The *bovine* dataset, provided by Rubinstein et al. [31], consisted of $5,984$ particle images. In all cases the particle images provided were 128×128, had a resolution of 2.8Å ($0.28nm$ per pixel) and CTF information for each particle image was provided. The noise level, σ, was estimated by computing the standard deviation of pixels around the boundary of the particle images.

To showcase the ability of our method to handle a dramatically different type of structure, a third dataset was synthesized by taking an existing structure from the Protein Data Bank[2], GroEL-GroES-(ADP)7 [39], and generating $40,000$ random projections according to the generative model. CTF, signal-to-noise level and other parameters were set realistically based on the *thermus* dataset values.

[2]Structure 1AON from http://pdb.org

This structure, as well as previously solved structures of the bovine and thermus ATP synthase molecules are shown in Figure 4. GroEL-GroES, was selected because it is structurally unlike either of the bovine or thermus ATP synthase molecules. Sample GroEL synthetic images can be see in Figure 11 (top left).

Results of our method on these datasets are shown in Figure 11. Sample particle images are shown, along with an iso-surface and slices of the final estimated density. Computing these reconstructions took less than 24 hours in all cases. Further, even at early iterations, reasonable structures are available. Figure 5 shows the estimated structure for the *thermus* dataset over time during optimization. Notably, after just one hour (during which only a fraction of the full dataset is seen), the low-resolution shape of the structure has already been determined.

Importance Sampling To validate our importance sampling approach we evaluated the error made in computing $\log p(\tilde{\mathcal{I}}|\theta, \tilde{\mathcal{V}})$ using IS against computing the exact sum in Equation (4) without IS. This error is plotted in Figure 6, along with the fraction of quadrature points used at various values of s_0. Based on these plots we selected a factor of $s_0 = 10$ for all experiments as a trade-off between accuracy and speed achieving a relative error of less then 0.1% while still providing significant speedups.

To see just how much of a speedup importance sampling provides in practice, we plotted in Figure 7 the fraction of quadrature points which needed to be evaluated during optimization. As can be seen initially, all quadrature points are evaluated but as optimization progresses and the density (and consequently the distribution over poses) becomes better determined importance sampling yields larger and larger speedups. At the full resolution, importance sampling provided more than a 100,000 fold speedup.

No prior knowledge of the orientation distribution was assumed. However, for many particles, certain views are

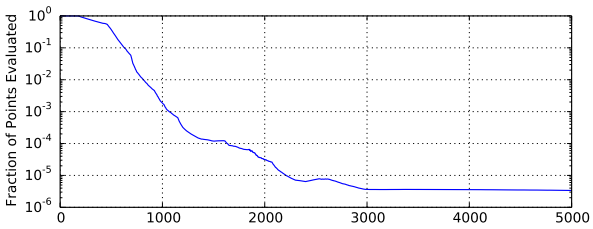


Figure 7: The fraction of quadrature points evaluated on average during optimization (horizontal axis is iterations). As resolution increases, the speedup obtained increases significantly yielding more than a 100,000 fold speedup.

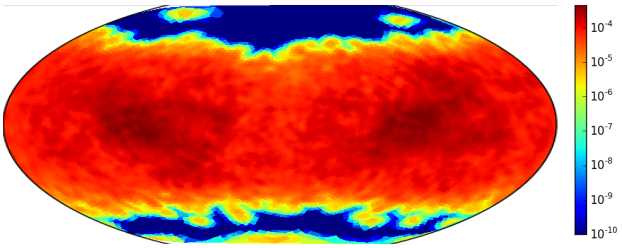
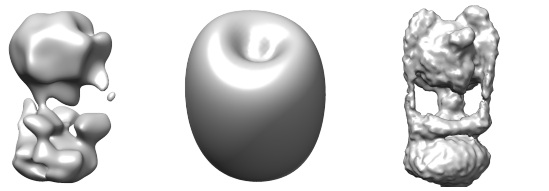


Figure 8: A Winkel-Tripel projection of the importance distribution of view directions, $\mathbf{q}^{\mathbf{R}}$, averaged over the *thermus* dataset at a typical iteration. Clearly visible is the equatorial belt of likely views, while axis aligned views (those on the top or bottom of the plot) are rarely seen.

more likely than others. This fact can be seen by examining the average importance distribution for the thermus dataset, shown in Figure 8 for a typical iteration. Here we can see clearly that the distribution of views forms an equatorial belt around the particle, while top or bottom views are rarely if ever seen. This phenomenon is well known for particles like these (*e.g.*, see [31] where this knowledge was used directly in estimation), validating our sampling approach and suggesting a use of this average importance distribution to supplement the uniform prior distribution in Eq. (14).

Initialization and Comparison to State-of-the-Art To compare this method to existing methods for structure determination, we selected two representative approaches. The first is a standard iterative projection matching scheme where images are matched to an initial density through a global cross-correlation search [11]. The density is then reconstructed based on these fixed orientations and this process is iterated. The second is the RELION package described in [34] which uses a similar marginalized model as our method but with a batch EM algorithm to perform optimization. We used publicly available code for both of these approaches on the *thermus* dataset and initialized using the density shown in Figure 5. We ran each method for a number of iterations roughly equivalent computationally



Projection Matching RELION Proposed Approach

Figure 9: Baseline comparisons to two existing standard methods. Iterative projection matching and reconstruction (left) and RELION [34] (middle). The proposed method (right) is able to determine the correct structure while projection matching and RELION both become trapped in poor local optima. See Fig. 9(middle) for comparison. All methods used the same random initialization shown in Fig. 5.

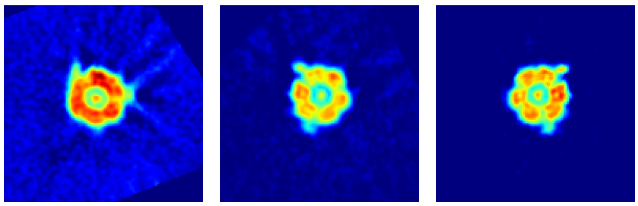


Figure 10: Slices through the reconstructions with (from left to right) uniform, CAR and exponential priors. The exponential prior does the best job of suppressing noise in the background without oversmoothing fine details within the structure. Blue corresponds to small or zero density and red corresponds to high density.

to the 5000 iterations used by our method and the results are shown in Figure 9. In both cases the approaches had clearly determined an incorrect structure and appeared to have converged to a local minimum as no further progress was made. Both projection matching and RELION have been used successfully for reconstruction by others and are not recommended to be used without a good initialization. Our results support this recommendation as neither approach converges from random initializations. In practice, it is difficult to construct good initializations for molecules of unknown shape [12], giving our proposed method a significant advantage.

Comparing Priors The above results used an exponential prior for the density at the voxels of $\mathcal{V}_i$, however the presented framework allows for any continuous and differentiable prior to be used. To demonstrate this, we explored two other priors: an improper (uniform) prior, $p(\mathcal{V}_i) \propto 1$, and a conditionally autoregressive (CAR) prior [5] $p(\mathcal{V}_i|\mathcal{V}_{-i}) = \mathcal{N}(\mathcal{V}_i|\frac{1}{26}\sum_{j\in\text{Nbhd}(i)}\mathcal{V}_j, \sigma^2_{CAR})$ which is a smoothness prior biasing each voxel towards the mean of its 26 immediate neighbours $\text{Nbhd}(i)$. Slices through the resulting densities on *thermus* under these priors are shown in Figure 10. With an improper uniform prior (Fig.

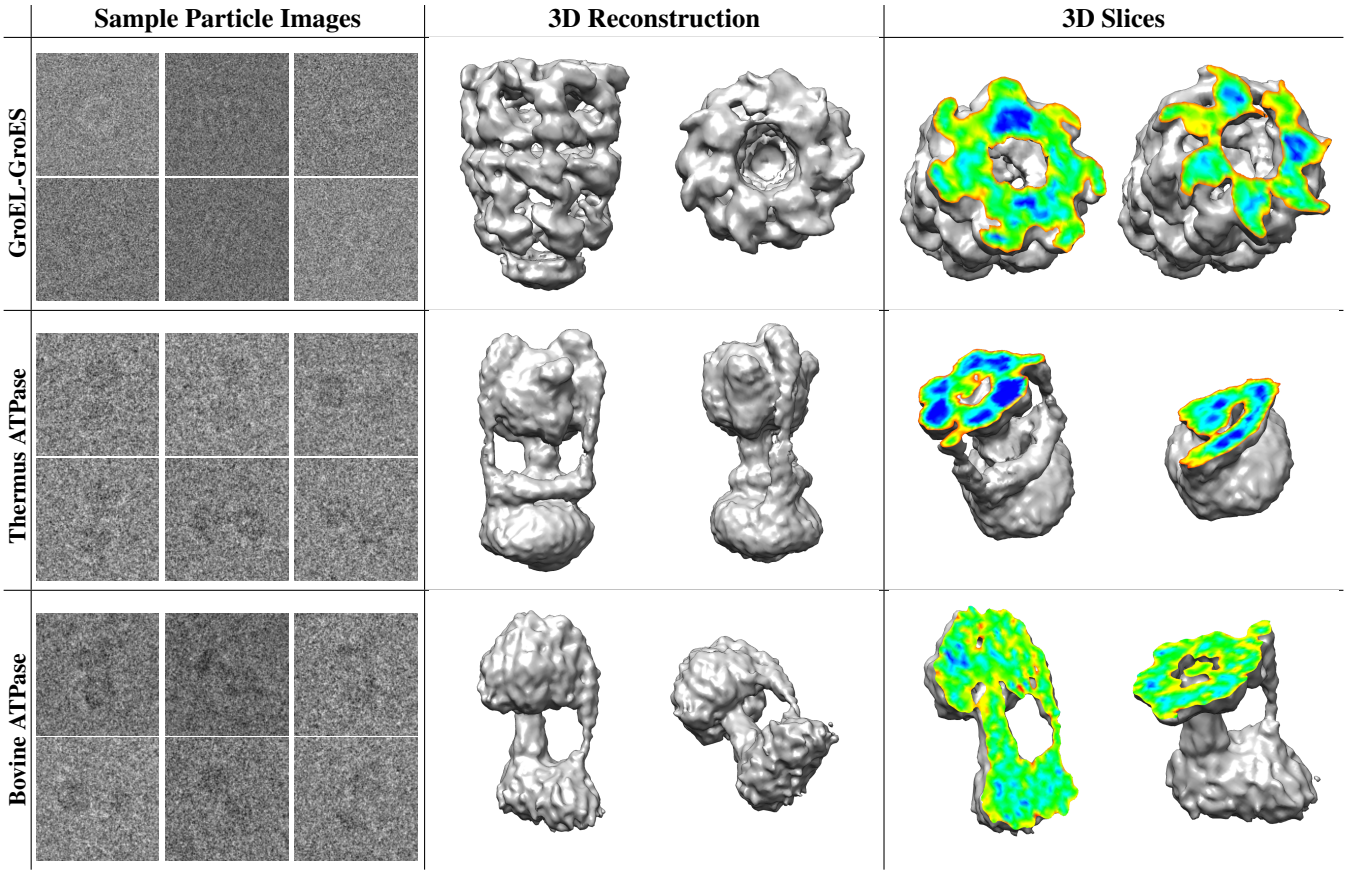


Figure 11: Sample particle images (left), an isosurface of the reconstructed 3D density (middle) and slices through the 3D density with colour indicating relative density (right) for GroEL-GroES (top), thermus thermophilus ATPase (middle) and bovine mitochondrial ATPase (bottom). The relative root expected mean squared error (RREMSE) on a held-out test set was 0.99, 0.96 and 0.98 with values relative to the estimated noise level. See Supplemental Material for more on the error measure. Reconstructions took a day or less on a 16 core workstation.

10,left), there is significant noise visible in the background. This noise is somewhat suppressed with the CAR prior (Fig. 10,middle) however the best results are clearly obtained using the exponential prior which suppresses the background noise without smoothing out internal details.

5. Conclusions

This paper introduces a framework for efficient 3D molecular reconstruction from Cryo-EM images. It comprises MAP estimation of 3D structure with a generative model, marginalization over 3D particle poses, and optimization using SAGD. A novel importance sampling scheme was used to reduce the computational cost of marginalization. The resulting approach can be applied to large stacks of Cryo-EM images, providing high resolution reconstructions in a day on a 16-core workstation.

The problem of density estimation for Cryo-EM is a fascinating vision problem. The low SNR in particle images makes it remarkable that any molecular structure can be estimated, let alone the high resolution densities which are now common. Recent research [23] suggests that the combination of new techniques and new sensors may facilitate atomic resolution reconstructions for arbitrary molecules. This development will be ground-breaking in both biological and medical research.

Beyond the work described in this paper, there remain a number of unresolved questions for future research. While an exponential prior was found to be effective, more sophisticated priors could be learned, potentially enabling higher resolution estimation without the need to collect more data and providing a kind of of atomic-scale super-resolution. The optimization problem is challenging, and, while SAGD was successful here, it is likely that more efficient stochastic optimization methods are possible by exploiting the problem structure to a greater degree. In order to encourage others to work on this problem, source code will be available from the authors' website.

Acknowledgements This work was supported in part by NSERC Canada and the CIFAR NCAP Program. MAB was funded in part by an NSERC Postdoctoral Fellowship. The authors would like thank John L. Rubinstein for providing data and invaluable feedback.

References

[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building rome in a day," in *ICCV*, 2009.

[2] S.-I. Amari, H. Park, and K. Fukumizu, "Adaptive method of realizing natural gradient learning for multilayer perceptrons," *Neural Computation*, vol. 12, no. 6, pp. 1399–1409, 2000.

[3] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.

[4] W. T. Baxter, R. A. Grassucci, H. Gao, and J. Frank, "Determination of signal-to-noise ratios and spectral snrs in cryo-em low-dose imaging of molecules," *J Struct Biol*, vol. 166, no. 2, pp. 126–32, May 2009.

[5] J. Besag, "Statistical analysis of non-lattice data," *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 24, no. 3, pp. 179–195, 1975.

[6] R. Bhotika, D. Fleet, and K. Kutulakos, "A probabilistic theory of occupancy and emptiness," in *ECCV*, 2002.

[7] J. de la Rosa-Trevín, J. Otòn, R. Marabini, A. Zaldívar, J. Vargas, J. Carazo, and C. Sorzano, "Xmipp 3.0: An improved software suite for image processing in electron microscopy," *Journal of Structural Biology*, vol. 184, no. 2, pp. 321–328, 2013.

[8] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, 2000.

[9] M. Gräf and D. Potts, "Sampling sets and quadrature formulae on the rotation group," *Numerical Functional Analysis and Optimization*, vol. 30, no. 7-8, pp. 665–688, 2009.

[10] J. Gregson, M. Krimerman, M. B. Hullin, and W. Heidrich, "Stochastic tomography and its applications in 3d imaging of mixing fluids," *ACM Trans. Graph. (Proc. SIGGRAPH 2012)*, vol. 31, no. 4, pp. 52:1–52:10, 2012.

[11] N. Grigorieff, "Frealign: high-resolution refinement of single particle structures," *J Struct Biol*, vol. 157, no. 1, p. 117–125, Jan 2007.

[12] R. Henderson, A. Sali, M. L. Baker, B. Carragher, B. Devkota, K. H. Downing, E. H. Egelman, Z. Feng, J. Frank, N. Grigorieff, W. Jiang, S. J. Ludtke, O. Medalia, P. A. Penczek, P. B. Rosenthal, M. G. Rossmann, M. F. Schmid, G. F. Schröder, A. C. Steven, D. L. Stokes, J. D. Westbrook, W. Wriggers, H. Yang, J. Young, H. M. Berman, W. Chiu, G. J. Kleywegt, and C. L. Lawson, "Outcome of the first electron microscopy validation task force meeting," *Structure*, vol. 20, no. 2, pp. 205 – 214, 2012.

[13] J. Hsieh, *Computed Tomography: Principles, Design, Artifacts, and Recent Advances*. SPIE, 2003.

[14] N. Jaitly, M. A. Brubaker, J. Rubinstein, and R. H. Lilien, "A Bayesian Method for 3-D Macromolecular Structure Inference using Class Average Images from Single Particle Electron Microscopy," *Bioinformatics*, vol. 26, pp. 2406–2415, 2010.

[15] J. Keeler, *Understanding NMR Spectroscopy*. Wiley, 2010.

[16] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," *IJCV*, vol. 38, no. 3, pp. 199–218, 2000.

[17] R. Langlois, J. Pallesen, J. T. Ash, D. N. Ho, J. L. Rubinstein, and J. Frank, "Automated particle picking for low-contrast macromolecules in cryo-electron microscopy," *Journal of Structural Biology*, vol. 186, no. 1, pp. 1 – 7, 2014.

[18] W. C. Y. Lau and J. L. Rubinstein, "Subnanometre-resolution structure of the intact Thermus thermophilus H+-driven ATP synthase," *Nature*, vol. 481, pp. 214–218, 2012.

[19] N. Le Roux and A. Fitzgibbon, "A fast natural Newton method," in *ICML*, 2010.

[20] N. Le Roux, P.-A. Manzagol, and Y. Bengio, "Topmoumoute online natural gradient algorithm," in *NIPS*, 2008, pp. 849–856.

[21] N. Le Roux, M. Schmidt, and F. Bach, "A stochastic gradient method with an exponential convergence rate for strongly convex optimization with finite training sets," in *NIPS*, 2012.

[22] V. J. Lebedev and D. N. Laikov, "A quadrature formula for the sphere of the 131st algebraic order of accuracy," *Doklady Mathematics*, vol. 59, no. 3, pp. 477 – 481, 1999.

[23] X. Li, P. Mooney, S. Zheng, C. R. Booth, M. B. Braunfeld, S. Gubbens, D. A. Agard, and Y. Cheng, "Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-em," *Nature Methods*, vol. 10, no. 6, pp. 584–590, 2013.

[24] S. P. Mallick, B. Carragher, C. S. Potter, and D. J. Kriegman, "Ace: Automated {CTF} estimation," *Ultramicroscopy*, vol. 104, no. 1, pp. 8–29, 2005.

[25] S. P. Mallick, S. Agarwal, D. J. Kriegman, S. J. Belongie, B. Carragher, and C. S. Potter, "Structure and view estimation for tomographic reconstruction: A bayesian approach," in *CVPR*, 2006.

[26] J. Martens, "Deep learning via hessian-free optimization," in *ICML*, 2010.

[27] J. A. Mindell and N. Grigorieff, "Accurate determination of local defocus and specimen tilt in electron microscopy," *Journal of Structural Biology*, vol. 142, no. 3, pp. 334–47, 2003.

[28] Y. Nesterov, "A method of solving a convex programming problem with convergence rate o (1/k2)," *Soviet Mathematics Doklady*, vol. 27, no. 2, pp. 372–376, 1983.

[29] B. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.

[30] L. Reimer and H. Kohl, *Transmission Electron Microscopy: Physics of Image Formation*. Springer, 2008.

[31] J. L. Rubinstein, J. E. Walker, and R. Henderson, "Structure of the mitochondrial atp synthase by electron cryomicroscopy," *The EMBO Journal*, vol. 22, no. 23, pp. 6182–6192, 2003.

[32] B. Rupp, *(2009). Biomolecular Crystallography: Principles, Practice and Application to Structural Biology*. Garland Science, 2009.

[33] S. H. W. Scheres, H. Gao, M. Valle, G. T. Herman, P. P. B. Eggermont, J. Frank, and J.-M. Carazo, "Disentangling conformational states of macromolecules in 3d-em through likelihood optimization," *Nature Methods*, vol. 4, pp. 27–29, 2007.

[34] S. H. Scheres, "RELION: Implementation of a Bayesian approach to cryo-EM structure determination ," *Journal of Structural Biology*, vol. 180, no. 3, pp. 519 – 530, 2012.

[35] F. Sigworth, "A maximum-likelihood approach to single-particle image refinement," *Journal of Structural Biology*, vol. 122, no. 3, pp. 328 – 339, 1998.

[36] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *ICML*, 2013.

[37] G. Tang, L. Peng, P. Baldwin, D. Mann, W. Jiang, I. Rees, and S. Ludtke, "EMAN2: an extensible image processing suite for electron microscopy," *Journal of Structural Biology*, vol. 157, no. 1, pp. 38–46, 2007.

[38] S. T. Tokdar and R. E. Kass, "Importance sampling: a review," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 1, pp. 54–60, 2010.

[39] Z. Xu, A. L. Horwich, and P. B. Sigler, "The crystal structure of the asymmetric GroEL-GroES-(ADP)7 chaperonin complex," *Nature*, vol. 388, pp. 741–750, 1997.

[40] J. Zhao, M. A. Brubaker, and J. L. Rubinstein, "TMaCS: A hybrid template matching and classification system for partially-automated particle selection," *Journal of Structural Biology*, vol. 181, no. 3, pp. 234 – 242, 2013.

A. Stochastic Optimization

This section provides algorithmic details of the Stochastic Averaged Gradient Descent (SAGD) optimization method used for MAP estimation. See the original SAGD paper [21] for details. Consider the objective function specified in Equation (6), rewritten as a sum of functions over subsets of the data:

$$\begin{aligned} f(\mathcal{V}) &= -\log p(\mathcal{V}) - \sum_{i=1}^{K} \log p(\tilde{\mathcal{I}}_i|\theta_i, \tilde{\mathcal{V}}) \\ &= \sum_{i=1}^{K} \left[-\frac{1}{K} \log p(\mathcal{V}) - \log p(\tilde{\mathcal{I}}_i|\theta_i, \tilde{\mathcal{V}}) \right] \\ &= \sum_{i=1}^{K} f_i(\mathcal{V}) \end{aligned}$$

At each iteration τ, SAGD computes the update given by

$$\mathcal{V}_{\tau+1} = \mathcal{V}_\tau - \frac{\epsilon}{L} \sum_{j=1}^{K} \left[d\mathcal{V}_j^\tau - \frac{1}{K} \frac{\partial}{\partial \mathcal{V}} \log p(\mathcal{V}) \right]$$

where $d\mathcal{V}_k^\tau$ is defined according to Equation (8). In practice, the sum in the above update equation is not computed at each iteration, but rather a running total is maintained and updated as follows:

$$\begin{aligned} \hat{\mathbf{g}}_\tau &= \sum_{k=1}^{K} d\mathcal{V}_k^\tau \\ \hat{\mathbf{g}}_{\tau+1} &= \hat{\mathbf{g}}_\tau - d\mathcal{V}_{k_\tau}^\tau + \mathbf{g}_{k_\tau}(\mathcal{V}_\tau) \end{aligned}$$

The SAGD algorithm requires a Lipschitz constant L which is not generally know. Instead it is estimated using a line search algorithm where an initial value of L is increased until the instantiated Lipschitz condition $f(\mathcal{V}) - f(\mathcal{V} - L^{-1}d\mathcal{V}) < \frac{\|d\mathcal{V}\|^2}{2L}$ is met. The line search for the Lipschitz constant L is only performed once every 20 iterations. Note that a more sophisticated line search could be performed if desired. A good initial value of L is found using a bisection search where the upper bound is the smallest L found so far to satisfy the condition and the lower bound is the largest L found so far which fails the condition. In between line searches, L is gradually decreased to try to take larger steps. The entire SAGD algorithm is provided in Algorithm (1).

B. Importance Sampling

Importance Sampling is a key part of the proposed reconstruction method for Cryo-EM and provides large speedups

Algorithm 1 SAGD

Initialize $\mathcal{V}$ and L
Initialize $\hat{\mathbf{g}} \leftarrow 0$
Initialize $d\mathcal{V}_k \leftarrow 0$ for all $k = 1..K$
for $\tau = 1..\tau_{\max}$ **do**
  Select data subset k_τ
  Compute objective gradient $\mathbf{g}_{k_\tau}(\mathcal{V})$
  $\hat{\mathbf{g}} \leftarrow \hat{\mathbf{g}} - d\mathcal{V}_{k_\tau} + \mathbf{g}_{k_\tau}(\mathcal{V})$
  $d\mathcal{V}_{k_\tau} \leftarrow \mathbf{g}_{k_\tau}(\mathcal{V})$
  $\mathcal{V} \leftarrow \mathcal{V} - \frac{\epsilon}{L}\left[\hat{\mathbf{g}} - \frac{\partial}{\partial \mathcal{V}} \log p(\mathcal{V})\right]$
  if mod(τ,20) == 0 **then**
    Perform line search
    while $f_{k_\tau}(\mathcal{V}) - f_{k_\tau}(\mathcal{V} - L^{-1} d\mathcal{V}_{k_\tau}) < \frac{\|d\mathcal{V}_{k_\tau}\|^2}{2L}$ **do**
      $L \leftarrow 2L$
    end while
  else
    $L \leftarrow \frac{K}{2^{\frac{1}{150}}}$
  end if
end for

Algorithm 2 Importance Sampling

Given ϕ_i for $i \in \mathfrak{I}$ from previous iteration
for $j \in 1..J$ **do**
  for $i \in \mathfrak{I}$ **do**
    Compute $\mathbf{K}_{i,j}$
  end for
end for
$\hat{\phi}_j \leftarrow \sum_{i \in \mathfrak{I}} \phi_i^{1/T} \mathbf{K}_{i,j} \quad \forall j \in 1..J$
$Z \leftarrow \sum_j \hat{\phi}_j$
$q_j \leftarrow (1-\alpha) Z^{-1} \hat{\phi}_j + \alpha \psi_j \quad \forall j \in 1..J$
$s \leftarrow \left(\sum_j q_j^2\right)^{-1}$
$N \leftarrow s_0 s$
$\mathfrak{I} \leftarrow \varnothing$
for $k \in 1..N$ **do**
  $i \leftarrow$ sample from q
  insert i into $\mathfrak{I}$
end for
Use $\mathfrak{I}$ to compute ϕ_i for next iteration

during optimization. We use importance sampling to efficiently compute the discrete sum in Equation (4). Note that importance sampling is applied independently for each image in the dataset, since the orientations and shifts which correspond to important terms in the discrete sum can be different for each image.

In practice, we split the outer sum in Equation (4) into a double summation, one over orientations on the sphere and one over in-plane rotations of images and projections. We then compute each of the three sums (over shift, in-plane rotation, and orientation) with and independent importance sampler. This is equivalent to computing the full sum in Equation (4) using a single importance sampler with an importance distribution that is factored into three parts, one for each of shift, in-plane rotation, and orientation. This factoring is necessary, as the memory requirements of storing a fully joint importance distribution for each image in the dataset would become infeasible for high-resolution reconstructions.

For each of the three importance samplers, the importance distribution at each iteration is constructed according to Equation (14). At the first iteration during which a particular image is seen, the importance distribution is simply uniform, and in fact we explicitly sample every point once. The ϕ values resulting from this computation are stored. At the next iteration during which the same image is seen, these ϕ values are used in Equation (14) to construct a non-uniform importance distribution which is then sampled from. We use a number of samples proportional to the effective sample size of the importance distribution, so the number of samples used naturally decreases as the importance distribution becomes more peaked, leading to large speedups at late iterations during optimization.

In Equation (14), the previous ϕ values are not directly used, but rather they are annealed by a temperature parameter and then smoothed by a kernel matrix. Both of these steps serve to guard against importance distributions which are too peaked around large ϕ values, which would inhibit the importance sampler from exploring the domain. The kernel matrix also serves the purpose of allowing use of ϕ values from a previous iteration even if the resolution of quadrature points being used has increased at the current iteration. The Von Mises-Fisher kernel is used for orientations and in-plane rotations, while a Gaussian kernel is used for shift:

$$\mathbf{K}_V(d_i, d_j; \kappa_V) \propto \exp(\kappa_V d_i^T d_j)$$
$$\mathbf{K}_G(t_i, t_j; \kappa_G) \propto \exp(-\kappa_G \|\mathbf{t}_i - \mathbf{t}_j\|^2)$$

where κ_V and κ_G are precision parameters for each kernel which are set based on the resolution of the quadrature scheme used at the previous ϕ values, d_i and d_j are the quadrature directions (in $\mathcal{S}^2$ for particle orientation and $\mathcal{S}^1$ for in-plane rotation, and $\mathbf{t}_i$ and $\mathbf{t}_j$ are the quadrature shift values (in $\mathbb{R}^2$).

The algorithm for constructing an importance distribution and sampling from it are given in Algorithm (2). The sampled values are then used to compute (12). Note that some quadrature points can end up being sampled multiple times, this is detected and the value reused to reduce computation.

C. Error Measures

Because ground-truth is rarely available for Cryo-EM, measuring accuracy is often difficult. Traditionally, the field has used the *Fourier Shell Correlation* (FSC) to measure

the resolution of a solved structure. The so-called gold-standard FSC works by splitting the dataset in half, estimating two densities separately and the computing the normalized correlation in Fourier space as a function of frequency. This curve would then be thresholded to provide an estimate of accuracy. However, we note that this measure is actually estimating the variance of the estimator, not the accuracy of the density it has produced. Further it is only theoretically justifiable when the estimator is unbiased, which is not true of the method proposed here or with other likelihood-based Bayesian methods such as RELION.

Instead, we introduce a novel metric based on reconstruction error of a held test set. To quantify the ability of marginal likelihood methods, such as ours, to model and explain the observed data we introduce the *Expected Mean Squared Error*

$$\mathcal{E}^2(\mathcal{I}|\theta,\mathcal{V}) \equiv E_{\mathbf{R},\mathbf{t}|\mathcal{I},\theta,\mathcal{V}}\left[\|\mathcal{I}-\mathbf{C}_\theta\mathbf{S}_\mathbf{t}\mathbf{P}_\mathbf{R}\mathcal{V}\|^2\right] \quad (15)$$

to be the expectation of the squared error between the image and its reconstruction under the image formation model. Note that the expectation is conditioned on the current density and the CTF parameters and is taken over the unknown pose and translation, $\mathbf{R}$ and $\mathbf{t}$. After switching to Fourier space and with some manipulation $\mathcal{E}^2(\mathcal{I}|\theta,\mathcal{V})$ becomes

$$Z^{-1}\int_{\mathbb{R}^2}\int_{\mathcal{SO}(3)}\|\tilde{\mathcal{I}}-\tilde{\mathbf{C}}_\theta\tilde{\mathbf{S}}_\mathbf{t}\tilde{\mathbf{P}}_\mathbf{R}\tilde{\mathcal{V}}\|^2 p(\tilde{\mathcal{I}}|\theta,\mathbf{R},\mathbf{t},\tilde{\mathcal{V}})p(\mathbf{R})p(\mathbf{t})d\mathbf{R}d\mathbf{t} \quad (16)$$

where the

$$Z=\int_{\mathbb{R}^2}\int_{\mathcal{SO}(3)} p(\tilde{\mathcal{I}}|\theta,\mathbf{R},\mathbf{t},\tilde{\mathcal{V}})p(\mathbf{R})p(\mathbf{t})d\mathbf{R}d\mathbf{t} \quad (17)$$

is a normalization constant. Computing this would be computationally expensive, instead we use an importance sampling based approximation, $\hat{\mathcal{E}}^2(\mathcal{I}|\theta,\mathcal{V})$,

$$\hat{Z}^{-1}\sum_{j\in\mathfrak{I}^\mathbf{R}}\sum_{\ell\in\mathfrak{I}^\mathbf{t}}\frac{w_j^\mathbf{R}w_\ell^\mathbf{t}}{N_\mathbf{R}q_j^\mathbf{R}N_\mathbf{t}q_\ell^\mathbf{t}}p_{j,\ell}\|\tilde{\mathcal{I}}-\tilde{\mathbf{C}}_\theta\tilde{\mathbf{S}}_\mathbf{t}\tilde{\mathbf{P}}_\mathbf{R}\tilde{\mathcal{V}}\|^2 \quad (18)$$

where

$$\hat{Z}=\sum_{j\in\mathfrak{I}^\mathbf{R}}\sum_{\ell\in\mathfrak{I}^\mathbf{t}}\frac{w_j^\mathbf{R}w_\ell^\mathbf{t}}{N_\mathbf{R}q_j^\mathbf{R}N_\mathbf{t}q_\ell^\mathbf{t}}p_{j,\ell} \quad (19)$$

is the approximation of the normalization constant. The above quantities can be readily computed along with the main likelihood computation using the same importance sampling scheme described above.

We compute the average value of $\hat{\mathcal{E}}^2(\mathcal{I}|\theta,\mathcal{V})$ on a held out set of test images whose gradients are never used. To normalize for different datasets we report the *Relative Root Expected Mean Squared Error* (RREMSE) as

$$\sqrt{\frac{1}{\sigma^2 N_\text{test}}\sum_\mathcal{I}\hat{\mathcal{E}}^2(\mathcal{I}|\theta,\mathcal{V})} \quad (20)$$

where the sum is taken over the test set which has N_test images and σ^2 is the noise variance of the dataset. Values near 1 indicate that the data is being well explained.